The Label-Translation Fallacy: A Four-Layer Validity Protocol for Multilingual Mental Health NLP

Wajdi Zaghouani
Northwestern University in Qatar
wajdi.zaghouani@northwestern.edu

Abstract

A familiar paper translates an English screener, fits a classifier to the new labels, reports a strong F_1, and claims to detect depression in another language community. Standard metrics can look fine while construct equivalence, instrument adaptation, and annotation ecology remain unshown. We call this the *label-translation fallacy*. Building on unified validity theory, measurement modeling, data statements, and cultural psychiatry, we propose a four-layer reporting protocol (construct, instrument, annotation ecology, and licensed claim) for authors and reviewers at venues such as MultiPsyche. We provide minimal evidence thresholds grounded in COSMIN and psychometric standards, concrete templates for L3 annotation reporting, and a decision tree for mapping evidence to claims. Three one-pass case studies show operationalizability. The contribution is integrative rather than experimental: we combine existing validity tools into an operational checklist so that multilingual mental health metrics are interpreted only as strongly as their constructs allow.

1 Introduction

Consider a stylized but familiar submission. An English PHQ-9 is translated into another language; posts or dialogues are labeled from the translated items or by an LLM judge; a model reports F_1 near 0.8; the abstract claims depression detection for that community. Nothing in the metric tells a reviewer whether the construct traveled, whether the cut-off still means the same thing, or whether the annotators shared the speakers' ontology. That is the failure case this paper is written to catch.

Foundation models make such pipelines easy: a prompt can be rewritten, a questionnaire machine-translated, an English social-media classifier applied elsewhere. The opportunity is real. So is the measurement risk. Surveys of mental health NLP find English-centric, depression-heavy resources and weak generalization even within English (Harrigian et al., 2021, 2020; Chancellor and De Choudhury, 2020). A multilingual survey of 108 datasets across 25 languages warns that disclosure norms change what models can observe (Bucur et al., 2026). Multilingual LLM evaluations further show that moving a task through machine translation, or judging non-English text with an English-centric model, changes measured performance in ways that are easy to mistake for clinical signal (Raihan et al., 2026; Xu and Abdullah, 2026).

A quieter problem sits underneath resource scarcity. Papers operationalize "depression," "anxiety," or "risk" by translating items from instruments such as the PHQ-9 and GAD-7 (Kroenke et al., 2001; Spitzer et al., 2006), by mapping local talk onto English diagnostic vocabularies, or by asking a model to judge severity in another language. Cross-cultural psychometrics does not automatically license that leap. A COSMIN review of the GAD-7 finds only nine eligible studies and describes the overall evidence for cross-cultural validity as very limited, even where a unitary factor structure is often recovered (Mantzari et al., 2026). Analyses of the PHQ-9 in eight low-resource antenatal samples find internal consistency without full measurement invariance: only partial metric invariance held, and only across four countries (Murray et al., 2022). Scores can be reliable and still mean different things. Cultural psychiatry has long treated idioms of distress as communicative systems rather than noisy translations of DSM criteria (Nichter, 1981; Kirmayer and Young, 1998; Lewis-Fernández et al., 2014).

This paper is written for MultiPsyche, the First Workshop on Multimodal, Multilingual, and Multicultural Mental Health and Psychotherapy, and for neighboring *ACL venues. We name the error the *label-translation fallacy* and offer a four-layer protocol as a diagnostic and reporting tool.

Proceedings of the 1st Workshop on Multimodal, Multilingual, and Multicultural Mental Health and Psychotherapy (MultiPsyche), pages 44–54
November 9, 2026 ©2026 Association for Computational Linguistics

Contributions. (1) We formulate the label-translation fallacy and locate it in unified validity theory and measurement modeling (Messick, 1995; Jacobs and Wallach, 2021). (2) We specify a four-layer protocol with minimal evidence thresholds grounded in COSMIN and classical psychometrics. (3) We provide a claim-licensing decision tree that operationalizes the mapping from L1-L3 evidence to L4 claims. (4) We provide concrete templates for L3 annotation ecology reporting. (5) We apply the layers to three published papers spanning instrument adaptation, synthetic multilingual supervision, and English social-media generalization, and walk through a hypothetical unlicensed case to show the protocol in use during review rather than only after publication. (6) We extend the layers to multimodal signals and show how the protocol relates to a longer lineage of reporting standards (PRISMA, data statements, datasheets, model cards, and TRIPOD+AI) without replacing any of them.

The protocol has not been tested for inter-reviewer reliability on a large corpus. Section 4 is a proof of operationalizability, not a systematic review.

2 Related Work

Validity as inference. Messick treats validity as the warrant for inferences from responses, including their social consequences. Content, criterion, and construct evidence are aspects of one argument, not interchangeable types. Jacobs and Wallach translate this into computational settings: many fairness harms follow from a mismatch between a theoretical construct and its operationalization. Mental health NLP is an acute case. Chancellor and De Choudhury already identified construct validity as a structural weakness of social-media prediction. What remains underspecified is how those critiques should change reporting when language and culture change together.

Instruments, COSMIN, and invariance. The PHQ-9 and GAD-7 are brief screeners validated primarily in Western clinical and primary-care settings (Kroenke et al., 2001; Spitzer et al., 2006; Kroenke et al., 2010). The COSMIN taxonomy distinguishes reliability, validity (including cross-cultural validity / measurement invariance), and responsiveness for patient-reported measures (Mokkink et al., 2010). Translation is not validation. Cognitive interviewing and cultural adaptation can require recasting statements as questions and adding visual response aids (Luitel et al., 2024). Measurement invariance testing asks whether the same latent variable is measured with the same structure across groups (Meredith, 1993). Configural invariance concerns the same factor pattern; metric (weak) invariance concerns equal loadings; scalar (strong) invariance concerns equal intercepts and is the usual prerequisite for comparing latent means across groups; disorder-prevalence estimation additionally requires the evidence described in Section 3. Partial invariance is a reason to justify pooling, not a footnote.

Mental health NLP resources. English and depression dominate available corpora (Harrigian et al., 2021). Models trained on one platform or time window often fail on another even without leaving English (Harrigian et al., 2020). Bucur et al. catalogue 108 non-English or multilingual social-media resources across 25 languages and emphasize cultural variation in disclosure (Bucur et al., 2026). Ethnic and cultural groups differ in how depression is lexicalized even in a shared language (Loveys et al., 2018).

Multilingual models and judges. Raihan et al. evaluate LLMs on mental health classification across several languages and their machine-translated counterparts (Raihan et al., 2026). Under the present protocol, those results speak to retrieval under a given operationalization; they do not, by themselves, license screening. Xu and Abdullah show that editing persona nationality and language, then asking LLM judges to score depression severity, yields inconsistent alignment across languages (Xu and Abdullah, 2026). Generator quality and judge quality are not cleanly separated; both are L2-L3 threats.

Culture is not a language code. Cross-cultural NLP distinguishes linguistic form from common ground, aboutness, and values (Hershcovich et al., 2022). DSM-5's Cultural Formulation Interview separates cultural syndromes, idioms of distress, and explanatory models (Lewis-Fernández et al., 2014). NLP annotation guidelines almost never make that distinction.

Reporting standards we extend. Data statements ask who produced the text and under what conditions (Bender and Friedman, 2018). Datasheets document motivation, composition, collection, and recommended uses of a dataset (Gebru

et al., 2021). Model cards document intended use and disaggregated evaluation of a trained system (Mitchell et al., 2019). TRIPOD+AI specifies reporting for clinical prediction models that use regression or machine learning (Collins et al., 2024). Annotator-level disagreement work shows that flattening labels can erase socially patterned perspectives (Prabhakaran et al., 2021). The present protocol adds *construct provenance*: who applied which mental health construct, with which instrument, in which community.

A longer lineage of reporting guidelines. Structured reporting requirements are not new to health research. PRISMA specifies what a systematic review must disclose so that a reader can judge the review's own validity, independent of its conclusions (Page et al., 2021). The common thread across PRISMA, TRIPOD+AI, datasheets, and model cards is that a checklist forces a claim to be traceable to a specific, inspectable decision rather than to the polish of the write-up. The four-layer protocol borrows this logic but targets a gap none of these standards close on their own: none of them asks whether a *psychological construct*, as opposed to a sampling frame or a modeling choice, survived a move across languages. TRIPOD+AI asks what population a model was validated on; it does not ask whether "depression" means the same thing in that population as in the source instrument. That question is what L1 and L2 add.

3 A Four-Layer Validity Protocol

Existing work explains construct validity, invariance, or cultural variation separately. Our contribution is to integrate them into a reporting and claim-licensing protocol. A label is a claim that should be justified at four layers, ordered by what each one presupposes: L1 fixes what psychological object is being talked about; L2 fixes how that object is turned into an observable instrument or proxy, which presupposes L1 being named; L3 fixes who actually produced the label using that instrument, which presupposes L2 being specified; and L4 states what the resulting numbers license someone to claim, which presupposes all three. Failure at an earlier layer cannot generally be resolved by a larger model at a later one, because a later layer can only inherit the validity of the layers it was built on.

L1: Construct. Name the psychological object in both the source definition and candidate local conceptualizations. Is the target major depressive disorder, nonspecific distress, grief, somatized suffering, expressions of suicidality or suicidal ideation, or help-seeking talk? These are not interchangeable. Local idioms should be considered as candidate constructs, not residual error (Nichter, 1981; Hinton and Lewis-Fernández, 2010).

L2: Instrument, Operationalization, and Minimal Evidence Thresholds. State how the construct is observed: translated PHQ/GAD item, hashtag, subreddit flair, clinician note, structured interview, or model-as-judge. Report the adaptation pathway (forward translation, back-translation, expert review, and cognitive interviewing, or a weaker subset). Cite invariance or differential item functioning (DIF) evidence for this language and population, or state that it is absent. Do not invent new cut-points.

The following thresholds are grounded in COSMIN (Mokkink et al., 2010) and Meredith's invariance taxonomy (Meredith, 1993):

- **Retrieval or ranking claims** ("classifier matches our proxy"): No invariance required. Report metric (F_1, AUROC).
- **Descriptive within-sample claims** ("sample X showed scores ≥ 10"): Report factor structure and internal consistency. Do not pool with other samples.
- **Compare structure or relationships across groups** (e.g., factor loadings, regression coefficients): Requires metric (weak) invariance. Per-group absolute fit can be read against Hu and Bentler's (Hu and Bentler, 1999) combination guideline (CFI near .95 with RMSEA near .06 or SRMR near .08) as contextual guidance, not a strict gate. Absolute fit is not evidence of invariance; invariance is evidenced by the *change* in fit between nested models. Following Chen (Chen, 2007), we define both deltas as deterioration rather than using one subtraction order for both, since CFI falls and RMSEA rises as fit worsens: $\Delta\text{CFI} = \text{CFI}_{\text{less-constr.}} - \text{CFI}_{\text{more-constr.}}$, and $\Delta\text{RMSEA} = \text{RMSEA}_{\text{more-constr.}} - \text{RMSEA}_{\text{less-constr.}}$ (the same order for both would let real RMSEA deterioration read as a negative, trivially passing, value). Invariance is supported when

ΔCFI $\leq$ .01 and ΔRMSEA $\leq$ .015 comparing the metric model to the configural one; these are contextual heuristics sensitive to sample size, group balance, and estimator, not fixed guarantees. If full metric invariance is rejected, partial metric invariance requires substantive justification for which items are freed, not merely better fit; document and exclude them from any structure-comparison claim. These thresholds assume continuous indicators; PHQ/GAD-type items are ordinal and need categorical invariance testing (configural/threshold/metric/scalar under a WLSMV-type estimator) with different procedures and cutoffs (Svetina et al., 2020).

- **Compare mean scores across groups**: Requires scalar (strong) invariance, not metric invariance. Metric invariance alone licenses comparisons of structure or relationships (above); it does not license comparing observed levels, because unequal item intercepts can shift a group's mean independently of any real difference in the underlying construct. Evaluate scalar invariance the same way, by nested-model comparison: constrain intercepts equal in addition to loadings, and compare fit to the metric model using ΔCFI $\leq$ .01 and ΔRMSEA $\leq$ .015 (Chen, 2007). If full scalar invariance is rejected, partial scalar invariance may be used only with substantive justification for which intercepts are freed; report them, and if no scalar evidence exists at all, do not report a mean difference as substantive. This licenses a claim about average scale scores only; it does not by itself license a claim about categorical disorder prevalence, which is a different claim handled below.

- **Transfer a cut-off across groups**: Requires two distinct pieces of evidence, not one. Scalar (strong) invariance establishes that scores are comparable; it does not establish that the transferred cut-off retains adequate diagnostic accuracy in the target group, nor make the proportion above it a prevalence estimate. Alongside invariance, report target-population calibration: diagnostic accuracy of the transferred cut-off against a clinical reference standard where one is available, or that this was not checked where one is not. This is the highest-stakes transfer case, since a biased or uncalibrated cut-off misclassifies individuals, not just averages. If either piece is missing, state this and do not report prevalence from the cut-off; uncalibrated transfer can inflate apparent prevalence several-fold even with a clinically validated source cut-off (Case 1, Section 4). If target-population accuracy is inadequate, *recalibration* is the remedy: re-estimate the cut-off against an appropriate clinical reference standard and validate it, preferably on independent data, rather than deriving a threshold from the target distribution alone, which cannot by itself establish clinical validity. Report the selection criterion, validation procedure, and uncertainty.

- **Derive or locally validate a cut-off (screening)**: A cut-off developed and validated *within* the target population against a clinical reference standard (e.g., a structured diagnostic interview) is a different claim from transfer and does not require cross-group invariance, because no cross-group equivalence is being claimed. It requires diagnostic-accuracy evidence instead: sensitivity, specificity, and ideally predictive values with the sample's base rate stated, since predictive values shift with prevalence. Report these separately for any subgroups (e.g., age bands) the cut-off is meant to cover (Case 1, Section 4, follows this route).

- **Estimate disorder prevalence**: Neither a locally derived nor a transferred cut-off licenses a prevalence claim on its own. A proportion above a cut-off is a screening-positive rate, not a prevalence estimate, unless accompanied by independent diagnostic ascertainment or an explicit misclassification correction (using sensitivity/specificity to adjust the raw rate); both routes additionally require a sampling design appropriate for population estimation, since neither fixes an unrepresentative sample. Report which route and sampling design were used, or state that the number is a screening-positive rate.

- **When invariance, calibration, or local validation is untested**: Acceptable if stated explicitly and L4 is narrowed. Do not hide gaps.

L3: Annotation Ecology and Positionality. Who applied the label, in which language, with

what codebook, and with what relation to the speakers? A minimum L3 report specifies: (1) annotator role and credentials (clinician, peer, crowdsourced, LLM); (2) language(s) of annotation; (3) codebook origin (DSM, ICD, local idioms, or mixed); (4) relation to speakers (embedded, clinical distance, lived experience); (5) inter-annotator agreement (Cohen's kappa or Fleiss kappa, reported with a confidence interval and the label base rates, since kappa is sensitive to prevalence and the same value means different things at different base rates; treat a conventional benchmark such as $\geq$ 0.60 as a prompt to inspect agreement further, not as an automatic pass, and investigate low-agreement items for ambiguity before concluding retraining is needed); (6) when annotator groups disagree (clinician vs. peer), report both label distributions instead of forcing a single gold label.

Disagreement is not automatically positionality. Before treating a clinician/lived-experience split as evidence about construct non-equivalence, rule out simpler explanations first, using standard reliability diagnostics: ambiguous item wording, codebook inconsistency, or rater fatigue can all produce the same disagreement pattern without any positionality effect at all. Report item-level agreement statistics and inspect items with low agreement for ambiguity before interpreting the split substantively. Once these are ruled out, disagreement that tracks annotator identity rather than item ambiguity is itself informative and should be documented, not collapsed into a single gold label.

L4: Licensed Claim. Separate the metric from the inference it supports. High F_1 on a translated proxy supports "this classifier retrieves texts matching proxy labels in Language X" only, not "detects depression in community C" or "readies for triage." The clinical stakes are concrete: a system strong at L4 but weak at L2-L3 may mis-rank people by cultural display rules and disclosure norms.

A decision tree (Figure 1) maps L1-L3 evidence to licensed L4 statements.

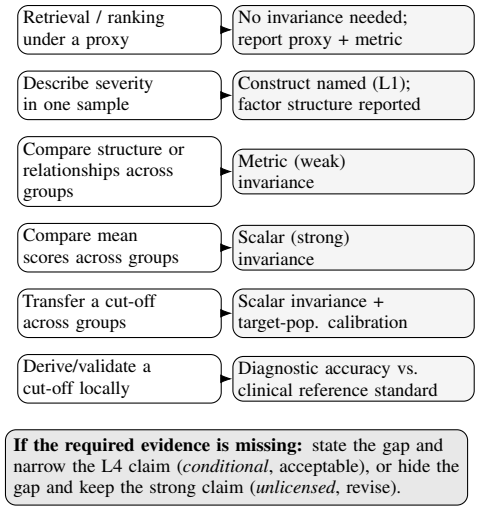


Figure 1: Evidence required at L1–L3 for six common L4 claim types. Comparing mean scores requires the stricter scalar invariance, not metric invariance; metric invariance alone licenses only structure/relationship comparisons. Transferring a cut-off across groups requires scalar invariance *and* target-population calibration, not invariance alone, since invariance establishes score comparability but not diagnostic accuracy at the cut-off; deriving and validating a cut-off within one population (Case 1, Section 4) requires diagnostic-accuracy evidence instead of invariance, since no cross-group claim is made. None of these claim types by itself licenses a disorder-prevalence estimate (Section 3). Table 1 gives the licensing map for the comparison and transfer claims above.

4 Three One-Pass Case Readings and Operationalizability

To show operationalizability, we read three published papers against the protocol. This is not an inter-rater reliability test or a systematic audit; it is a proof that the protocol can be applied to real work. We anchor each layer assignment to what the paper explicitly states wherever possible, but assigning a statement to L1–L4 and reaching a verdict is our interpretation, not a transcription; a different reader could draw the boundaries differently, which is what the inter-reviewer reliability study in Section 9 would measure.

Case 1: Instrument adaptation (Luitel et al., 2024). This paper adapts and validates the PHQ-A and GAD-7 for adolescents in Nepal. L1: Source constructs (adolescent depression and generalized anxiety, per DSM-5) are mapped to local Nepali conceptualizations through a qualitative phase: ten focus group discussions with 56 participants and 25 cognitive interviews, conducted in Nagarjun Municipality in July–August 2021. L2: The adaptation pathway (translation, expert review, and the qualitative work above) is reported in detail; the validation step, however, is diagnostic accuracy against a clinical gold standard, not configural/metric/scalar invariance testing. In a Phase 2 validation sample (Kathmandu Metropolitan City and Budhanilkantha and Tokha Municipalities), screening scores are compared against the Kiddie Schedule for Affec-

Evidence pattern	Licensed L4 statement
Proxy labels, no L2 adaptation	Retrieval under this operationalization only
Adapted instrument, configural + metric, no scalar	Comparison of structure/loadings; no prevalence
Scalar invariance + L3 documented	Cautious means comparison; still not diagnostic
Locally derived cut-off, diagnostic accuracy vs. clinical standard reported	Screening claim within the validated population only; no cross-population claim
Transferred cut-off, scalar invariance + target-population calibration reported	Screening claim in target population; prevalence additionally requires appropriate sampling and diagnostic ascertainment or justified misclassification correction
Gap at L2 or L3, explicitly stated and L4 narrowed	Acceptable (transparency is standard)
Gap hidden, L4 overclaimed	Not acceptable; revise

Table 1: Claim-licensing map.

tive Disorders and Schizophrenia (K-SADS) structured clinical interview. The paper reports age-stratified cut-offs: for younger adolescents (12–14 years), a PHQ-A cut-off of $\geq$ 13 reaches sensitivity 0.93 and specificity 0.80 (PPV 0.33, NPV 0.99); a GAD-7 cut-off of ≥ 8 reaches sensitivity 0.70 for both younger and older adolescents, with specificity 0.67 (younger) and 0.71 (older). L3: Diagnostic status comes from the K-SADS interview administered by trained personnel rather than from independent community annotators, so this case does not exercise the clinician/peer-annotator agreement pattern Section 3 otherwise asks for; the paper does not report an inter-annotator kappa in that sense. L4: The paper narrows its claims accordingly: it reports screening properties for this validated sample (fair for depression; poor for anxiety, with high false-positive rates) and states explicitly that even these clinically validated cut-offs would inflate population-level prevalence estimates by 2–4 fold if applied without adjustment.

Protocol verdict: Compliant. This paper's L2 evidence is diagnostic accuracy against a clinical interview (K-SADS), the "derive or locally validate a cut-off" route in Section 3, not the measurement-invariance route used for cross-group transfer; the two routes require different evidence because they license different claims. The paper's L4 statements, screening properties plus an explicit prevalence-inflation warning, claim no more than this evidence supports, and it does not report unadjusted-cut-off prevalence as a population estimate or claim cross-population transfer. The qualitative phase and K-SADS design are in the paper's Methods; Table 4 reports PHQ-A diagnostic accuracy against K-SADS and Table 5 reports GAD-7 diagnostic accuracy against K-SADS, each by age group; the prevalence-inflation warning is in its Discussion. For full sampling detail, consult Luitel et al. (2024) directly.

Case 2: Synthetic multilingual supervision (Xu and Abdullah, 2026). This paper generates English clinical personas, reuses them with language changed, and has LLMs judge severity in Mandarin, Bengali, and Hindi. L1: Assumes portable severity construct. L2: Operationalization is generation + LLM judge, not adapted instrument. Invariance testing is not conducted (and not relevant to generation). L3: Annotators are LLM judges; language-linked performance drops are measurement results, not noise. L4: The paper itself is appropriately hedged, reporting that judge models show inaccuracies in non-English severity assessment and that adding nationality/language parameters to a persona "might not be adequate" because it can introduce clinical inconsistency across languages; it frames this as a systemic limitation of English-centric personas rather than asserting that the severity construct categorically fails to transfer (Section 4.1.1 and Table 2 report the accuracy gap across languages directly). Read at L4, this raises a construct-transfer concern rather than settling it: construct equivalence across the three target languages remains unestablished, not disproven.

Protocol verdict: Compliant. The paper does not overclaim; it acknowledges measurement fragility.

Case 3: English generalization (Harrigian et al., 2020). Trains social-media mental health models in one setting and tests in others, across datasets whose proxy-label construction is summarized in the paper's Table 1 (self-report phrase matching, subreddit-based controls, or community participation, depending on the dataset). L1-L2: Constructs and proxies vary by dataset. L3: Labels derive from dataset-specific proxies, including self-reported diagnoses and community participation, not a single uniform mechanism. L4: Section 5.1 and Table 2 report that in-domain metrics do not license out-of-domain detection; cross-dataset F_1 drops substantially when a model trained on one dataset's proxies is evaluated on another's. If English-to-English generalization already breaks under this proxy diversity, multilingual and cross-cultural extension needs stricter validation.

Protocol verdict: Compliant. The paper documents where generalization fails.

All three papers pass the protocol test as we read them: in these illustrative readings, we identified no claim exceeding the evidence summarized above. Three papers, chosen as described below rather than sampled at random, cannot establish that this holds of the field generally.

Why these three, and what they show together. The three papers were not selected because they are exemplary; they were selected because each sits at a different point in the protocol's L2 axis: adapted instrument validated against a clinical diagnostic interview (Luitel), generation-plus-judge with no instrument at all (Xu and Abdullah), and same-language proxy labels that still fail to generalize (Harrigian). All three land on the same side of L4: none claims more than its L1–L3 evidence supports. That convergence is the point. A protocol that only worked on instrument-adaptation papers, or only on LLM-judge papers, would not be a general reporting tool. Reading these three together also shows that L4 discipline is not a proxy for methodological difficulty: Xu and Abdullah's operationalization is weaker at L2 than Luitel's, yet their L4 claim is just as appropriately narrow, because narrowing a claim costs nothing but honesty. The harder failure mode the protocol is built to catch (strong L4 language riding on an untested L2) did not appear in any of the three, which is itself a limitation of this sample rather than evidence that the failure mode is rare; overclaiming papers are, almost by construction, less likely to state their gaps clearly enough for a one-pass reading to reconstruct them, which is a form of survivorship bias worth naming rather than hiding.

5 Applying the Protocol: A Worked Walkthrough

The case readings in Section 4 show the protocol applied to finished papers. This section walks through the reasoning a reviewer or author would do *while* deciding what a result licenses, using a hypothetical submission. The scenario below is invented for illustration; it is not a description of any real paper, and no numbers in it should be read as data.

Scenario. A submission trains a classifier on Reddit posts labeled "depressed" or "not depressed" by prompting an LLM with the English PHQ-9 items translated into Swahili via machine translation. The classifier reaches a strong F_1 on a held-out split. The abstract states that the system "detects depression in Swahili-speaking social media users."

Step 1 (L1). Is the construct named in both source and local terms? The paper defines the target only as "depression," with no discussion of how depression or distress is talked about in Swahili-speaking communities, and no consideration of local idioms as candidate constructs in their own right. This is a gap, and it is not stated as one.

Step 2 (L2). What is the operationalization, and what adaptation evidence exists? The instrument is a machine-translated PHQ-9 used only to prompt an LLM judge, not administered to speakers or reviewed by a bilingual clinician. No forward/back-translation, no cognitive interviews, no invariance testing of any kind is reported. Under the thresholds in Section 3, this operationalization supports a retrieval claim and nothing stronger.

Step 3 (L3). Who applied the label? The LLM judge is the only annotator; there is no human agreement statistic, no codebook beyond the prompt itself, and no relationship between the judge and the speaker community to describe. This is again a gap, unstated.

Step 4 (L4, via Figure 1). The abstract's claim is "detects depression in Swahili-speaking users," which reads as a description/screening-level claim. A screening claim requires either validation against an appropriate clinical reference standard in the target population (if a cut-off is derived locally), or scalar invariance evidence together with target-population calibration (if a cut-off is transferred from elsewhere); an adapted instrument is a prerequisite for either route. The construct is named (Step 1) but not adequately defined or culturally contextualized, and the operationalization lacks adaptation evidence and validation against a clinical reference standard. The description/screening claim is **unlicensed**; the evidence actually collected supports only a retrieval claim.

Revision path. The same result can be reported honestly without new experiments: "Our classifier predicts LLM-assigned proxy labels generated using machine-translated PHQ-9 items, achieving [metric] on a held-out split. Clinical validity in Swahili-speaking communities remains unestablished." Nothing about the model changes; what changes is what the authors say the number means. This is the entire function of the protocol: it does not ask for more experiments, only for the claim

to stop outrunning the evidence that was actually collected.

6 L3 Annotation Ecology: Concrete Template

To help authors report L3 clearly:
Annotator Role: State explicitly: clinician (name specialty, years experience), community health worker (training pathway), crowdsourced (platform, qualification requirements), peer with lived experience (specify which condition), or LLM (model name, version, prompt).
Language(s) of Annotation: Which language(s) were inputs presented in? Are annotators native speakers of that language? If code-switching, was it included in training/examples?
Codebook / Training: Was a codebook used? If yes, cite its origin: DSM-5 diagnostic criteria, ICD-11, local idioms from qualitative interviews, or mixed. How many hours of training? Did training include examples from the target language/culture?
Relation to Speaker Community: Are annotators from the speaker community? Do they have lived experience with the condition? Or are they clinical staff (if so, what geographic/cultural distance)? This affects what kinds of suffering or distress they recognize.
Agreement & Disagreement: Report Cohen's kappa (or Fleiss for 3+) with a 95% CI and the label base rates. Treat κ as a prompt for further inspection rather than an automatic gate: state the adjudication rule used regardless of the value, and investigate items with low agreement for ambiguity before attributing it to annotator disagreement. If clinician and peer labels disagreed: report both distributions, do not force a single label.
Sensitive Point: A disagreement pattern that tracks annotator positionality (clinical distance vs. lived experience) is worth reporting, not hiding, but only once the simpler explanations in Section 3 (ambiguous items, codebook inconsistency, rater fatigue) have been checked and ruled out. It may inform questions about construct interpretation, after alternative explanations for disagreement have been examined; it does not establish construct non-equivalence on its own.

7 Extending the Protocol to Multimodal Signals

MultiPsyche's scope is explicitly multimodal as well as multilingual, and the four layers apply, with one addition, when speech or video are involved. Cultural *display rules* are learned norms governing who may show which emotion to whom, and they vary systematically across cultures rather than reflecting a universal mapping from internal state to expression (Ekman and Friesen, 1969; Matsumoto, 1990). A facial-affect or prosody model trained in one display-rule regime can therefore misread the same underlying distress in another: a flattened affect that reads as "masking" in one culture's display norms may simply be the expected register for discussing hardship in another. This is an L1–L2 problem in a different modality, not a new failure mode: the construct (what counts as an expression of distress) and the instrument (which signal is treated as evidence of it) both need the same local-conceptualization check that Section 3 asks for text.

Interview corpora that combine audio-visual features with questionnaire-derived labels, such as the DAIC family, illustrate why the L2 report needs a modality row (Gratch et al., 2014): a corpus collected in one clinical and cultural setting carries that setting's display-rule regime, and a model trained on it inherits the regime along with the signal. Missing or degraded channels (a camera angle that clips facial affect, a phone microphone that flattens prosody) belong in the same report, since a model silently reweighting toward whichever channel survives will do so unevenly across recording conditions that themselves may correlate with clinical setting or region. We do not extend the protocol's evidence thresholds (Section 3) to specific multimodal fusion methods; that is beyond the scope of a text-first protocol. We do ask that a multimodal submission fill one additional row: which signals were used, which display-rule regime the training data reflects, and which channels were missing or degraded, and for whom.

A further restriction applies to psychotherapeutic dialogue specifically, which MultiPsyche also covers. A model that retrieves turns that sound supportive is not thereby a therapist, and alliance, rupture, and risk disclosure are interactional achievements between two parties, not properties of a single utterance that a classifier can score in isolation. Where a session involves language mediation (an interpreter, or a system translating between patient and clinician in real time), that mediation is a clinical act with its own validity burden, not a preprocessing step that can be assumed lossless.

8 Implications for MultiPsyche

MultiPsyche asks for multimodal, multilingual, and multicultural work on assessment and psychotherapy. The protocol is a way to keep those axes from collapsing into "we translated the benchmark."

First, multilingual results are not settled by model scale. Translation is a hypothesis about meaning.

Second, synthetic therapy data and LLM-as-judge pipelines are clinical artifacts with a validity burden.

Third, multicultural evaluation is not only a fairness check after an English system works.

Where to put the protocol report: In dataset/methods section alongside data statements and datasheets. The licensed-claim sentence belongs in abstract and limitations. Model cards should restate L4 so downstream users do not promote a retrieval model into a screener.

9 Limitations and Future Work

Peer review raised three concerns that this camera-ready revision narrows or corrects (Section 3, Section 4) but does not fully resolve, and we state them directly rather than only in the sections they touch.

First, reviewers noted that the protocol has not been empirically validated for inter-reviewer reliability. Section 4 is a three-paper proof of concept, not an inter-rater reliability study or systematic audit. The protocol should be tested by having 10-20 reviewers independently apply it to the same papers and measuring agreement. We have not conducted this study.

Second, reviewers noted that the evidence thresholds are borrowed rather than validated for this setting. The protocol was not reviewed by clinicians, psychometricians, or multilingual community experts. Thresholds (e.g., the Hu and Bentler CFI/RMSEA fit guidelines, the $\Delta\text{CFI} \leq .01$ invariance criterion, and conventional kappa benchmarks) are standard in psychometrics and borrowed from existing standards, but their suitability specifically for multilingual mental health NLP, and for ordinal screening items in particular, has not been separately validated.

Third, reviewers questioned whether three case studies support a claim of general applicability, and asked us to clarify what is novel given that each individual tool (invariance testing, data statements, annotation reporting) already exists elsewhere. We agree that three cases do not establish generalizability; they were chosen to span different points on the protocol's L2 axis (Section 4, final paragraph) rather than to sample the field, and a claim that the protocol works broadly should wait for the inter-reviewer reliability study above. The contribution is integrative rather than inventing a new individual tool: to our knowledge, no prior work combines construct validity, measurement invariance, and annotation-ecology reporting into one claim-licensing procedure for this literature, and that integration, rather than any single threshold, is what we are asking authors and reviewers to adopt.

We additionally assume "culture" can be operationalized within language boundaries, which is often false. Culture is not nationality or language code.

10 Ethical Considerations

Better measurement can still enable surveillance; a protocol that makes labels more defensible does not make the underlying act of labeling distress talk automatically benign. Licensed-claim discipline is one ethical control among several that are needed: a system should not inherit clinical authority from a proxy label simply because that label now comes with a validity report attached. Annotation guidelines that overwrite local idioms with DSM categories risk pathologizing culturally normative grief, religiosity, or family obligation, and doing so under the banner of a more rigorous protocol does not remove the harm; it only makes the harm better documented.

The protocol also has a distributional cost that is worth naming rather than assuming away. Meeting L2's evidence thresholds (cognitive interviews, invariance testing, clinician-hours for annotation) requires resources that are unevenly available, and a strict reading of the thresholds could in practice penalize exactly the low-resource-language research this paper is meant to support. We have tried to guard against this by making "untested, explicitly stated" an acceptable L2 report rather than a disqualifying one (Section 3), but a checklist can still be used as a gate by a reviewer who reads absence of invariance testing as absence of rigor rather than as an honestly reported constraint. Whether the protocol is used this way in practice is an empirical question about reviewing culture, not something this paper's design can settle on its own.

Finally, the protocol does not by itself decide who should have standing to adjudicate a construct

dispute: a clinician trained in one diagnostic tradition, or a lived-experience annotator whose community uses a different idiom for the same distress. We have deliberately not resolved this in the protocol (Section 3, L3), beyond requiring that both perspectives be reported when they disagree rather than collapsed into one. That is a narrower ethical commitment than resolving the underlying question of authority, and we do not want the presence of an L3 report to be mistaken for having resolved it. Researchers working with non-English or non-Western clinical language should budget for community expertise as a first-class cost of the study, not as a preprocessing step delegated to translation. Dual use cannot be eliminated by a reporting protocol; it can be narrowed by refusing to let a retrieval result be described as a diagnostic one, and honest gaps remain preferable to fabricated precision at every layer of this protocol.

References

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Ana-Maria Bucur, Marcos Zampieri, Tharindu Ranasinghe, and Fabio Crestani. 2026. A survey on multilingual mental disorders detection from social media data. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 893–918, Rabat, Morocco. Association for Computational Linguistics.

Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: A critical review. *npj Digital Medicine*, 3:43.

Fang Fang Chen. 2007. Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, 14(3):464–504.

Gary S. Collins, Karel G. M. Moons, Paula Dhiman, Richard D. Riley, Andrew L. Beam, Ben Van Calster, Marzyeh Ghassemi, Xiaoxuan Liu, Johannes B. Reitsma, Maarten van Smeden, et al. 2024. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385:e078378.

Paul Ekman and Wallace V. Friesen. 1969. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, 1(1):49–98.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM*, 64(12):86–92.

Jonathan Gratch, Ron Artstein, Gale M. Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Albert Rizzo, and Louis-Philippe Morency. 2014. The distress analysis interview corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, pages 3123–3128, Reykjavik, Iceland.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2020. Do models of mental health based on social media data generalize? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3774–3788. Association for Computational Linguistics.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2021. On the state of social media data for mental health research. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology*, pages 15–24. Association for Computational Linguistics.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013. Association for Computational Linguistics.

Devon E. Hinton and Roberto Lewis-Fernández. 2010. Idioms of distress among trauma survivors: Subtypes and clinical utility. *Culture, Medicine, and Psychiatry*, 34(2):209–218.

Li-tze Hu and Peter M. Bentler. 1999. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6(1):1–55.

Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 375–385.

Laurence J. Kirmayer and Allan Young. 1998. Culture and somatization: Clinical, epidemiological, and ethnographic perspectives. *Psychosomatic Medicine*, 60(4):420–430.

Kurt Kroenke, Robert L. Spitzer, and Janet B. W. Williams. 2001. The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, 16(9):606–613.

Kurt Kroenke, Robert L. Spitzer, Janet B. W. Williams, and Bernd Löwe. 2010. The patient health questionnaire somatic, anxiety, and depressive symptom

scales: A systematic review. *General Hospital Psychiatry*, 32(4):345–359.

Roberto Lewis-Fernández, Neil Krishan Aggarwal, Sofie Bäärnhielm, Hans Rohlof, Laurence J. Kirmayer, Mitchell G. Weiss, Sushrut Jadhav, Ladson Hinton, Renato D. Alarcón, Dinesh Bhugra, et al. 2014. Culture and psychiatric evaluation: Operationalizing cultural formulation for DSM-5. *Psychiatry: Interpersonal and Biological Processes*, 77(2):130–154.

Kate Loveys, Jonathan Torrez, Alex Fine, Glen Moriarty, and Glen Coppersmith. 2018. Cross-cultural differences in language markers of depression online. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology*, pages 78–87, New Orleans, LA. Association for Computational Linguistics.

Nagendra P. Luitel, Damodar Rimal, Georgia Eleftheriou, Kelly Rose-Clarke, Suvash Nayaju, Kamal Gautam, Sagun Ballav Pant, Narmada Devkota, Shruti Rana, Jug Maya Chaudhary, Bhupendra Singh Gurung, Jill Witney Åhs, Liliana Carvajal-Velez, and Brandon A. Kohrt. 2024. Translation, cultural adaptation and validation of patient health questionnaire and generalized anxiety disorder among adolescents in Nepal. *Child and Adolescent Psychiatry and Mental Health*, 18:74.

Anastasia Mantzari, Lorena Botella-Juan, Prateek Yadav, Grace Lavelle, Ana Henriques, Daniel Munblit, Amy Ronaldson, Alex Dregan, Ioannis Bakolis, Mariana Pinto da Costa, Maria Carmen Viana, Alba Marcos-Delgado, Antonio J. Molina, Vicente Martín, Jose M. Valderas, Gemma Vilagut, Jordi Alonso, and Jorge Arias De La Torre. 2026. Cross-cultural validity of the generalized anxiety disorder 7-item scale (GAD-7): A systematic review and standardized assessment. *Epidemiology and Psychiatric Sciences*, 35:e24.

David Matsumoto. 1990. Cultural similarities and differences in display rules. *Motivation and Emotion*, 14:195–214.

William Meredith. 1993. Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4):525–543.

Samuel Messick. 1995. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9):741–749.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229.

Lidwine B. Mokkink, Caroline B. Terwee, Donald L. Patrick, Jordi Alonso, Paul W. Stratford, Dirk L. Knol, Lex M. Bouter, and Henrica C. W. de Vet. 2010. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology*, 63(7):737–745.

Aja Louise Murray, Chad Lance Hemady, Huyen Do, Michael Dunne, Sarah Foley, Joseph Osafo, Siham Sikander, Bernadette Madrid, Adriana Baban, Diana Taut, Catherine L. Ward, Asvini Fernando, Vo Van Thang, Manuel Eisner, Claire Hughes, Pasco Fearon, Sara Valdebenito, Mark Tomlinson, Arunasalam Pathmeswaran, and Susan Walker. 2022. Measuring antenatal depressive symptoms across the world: A validation and cross-country invariance analysis of the patient health questionnaire-9 (PHQ-9) in eight diverse low-resource settings. *Psychological Assessment*, 34(11):993–1007.

Mark Nichter. 1981. Idioms of distress: Alternatives in the expression of psychosocial distress: A case study from South India. *Culture, Medicine and Psychiatry*, 5(4):379–408.

Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, et al. 2021. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372:n71.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. In *Proceedings of the Joint 15th Linguistic Annotation Workshop and 3rd Designing Meaning Representations Workshop*, pages 133–138. Association for Computational Linguistics.

Nishat Raihan, Sadiya Sayara Chowdhury Puspo, Ana-Maria Bucur, Stevie Chancellor, and Marcos Zampieri. 2026. Large language models for mental health: A multilingual evaluation. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*, pages 335–346, Rabat, Morocco. Association for Computational Linguistics.

Robert L. Spitzer, Kurt Kroenke, Janet B. W. Williams, and Bernd Löwe. 2006. A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine*, 166(10):1092–1097.

Dubravka Svetina, Leslie Rutkowski, and David Rutkowski. 2020. Multiple-group invariance with categorical outcomes using updated guidelines: An illustration using mplus and the lavaan/semtools packages. *Structural Equation Modeling: A Multidisciplinary Journal*, 27(1):111–130.

Yunkai Xu and Saeed Abdullah. 2026. Creating multilingual mental health dialogue datasets: Limits of persona-based localization via nationality and language. In *Proceedings of the 11th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2026)*, pages 138–152, San Diego, California, USA. Association for Computational Linguistics.